Improving language identification of web page using optimum profile

Abstract

Language is an indispensable tool for human communication, and presently, the language that dominates the Internet is English. Language identification is the process of determining a predetermined language automatically from a given content (e.g., English, Malay, Danish, Estonian, Czech, Slovak, etc.). The ability to identify other languages in relation to English is highly desirable. It is the goal of this research to improve the method used to achieve this end. Three methods have been studied in this research are distance measurement, Boolean method, and the proposed method, namely, optimum profile. From the initial experiments, we have found that, distance measurement and Boolean method is not reliable in the European web page identification. Therefore, we propose optimum profile which is using N-grams frequency and N-grams position to do web page language identification. The result show that the proposed method gives the highest performance with accuracy 91.52%.